Automated Cyberbullying Detection

Motivation and Problem Definition:

Social Network websites becomes the major environment for the users to communicate and express their opinions. But, there are some users use the social network websites in negative way like cyberbullying.

The united states government defines the Cyberbullying as a bullying that takes place over digital devices like cell phones, computers, and tablets. And, it can occur through SMS, Text, and apps, or online in social media, forums, or gaming where people can view, participate in, or share content. Cyberbullying includes sending, posting, or sharing negative, harmful, false, or mean content about someone else. It can include sharing personal or private information about someone else causing embarrassment or humiliation [1]. Moreover, multiple studies mention that cyberbullying can affect any anyone especially young people and they are at high risk. For example, National Center for Education Statistics and Bureau of Justice Statistics reported that 9% of students in grades 6–12 experienced cyberbullying [2]. Moreover, Kosciw, J. G. et al. say that 55.2% of LGBTQ students experienced cyberbullying [3].

Cyberbullying has negative consequences on victim. For example, Sourander et al. say that cyberbullying leads to serious pathological experience such as depression, self-harm and suicide attempt [4]. In addition, S. Hinduja and J. W. Patchin say that the effects of cyberbullying can start from temporary anxiety to suicide [5]. Kowalski and limber mention that 90% of the young people victims don't tell their parent about their cyberbullying experience [6].

Manual detection of cyberbullying is very difficult because the huge volume of data on social network websites. Therefore, accurate and automated detection of cyberbullying is more effective. Several studies focused mainly on the content-based feature such as profanity, pronouns, bags of word, term frequency inverse document (TFIDF) and cyberbullying words. For example, Foong and Oussalah present an automated cyberbullying system detection. The system is based on natural language processing, text mining and machine learning to detect cyberbullying. The authors employed different textual features for the classifier such TF-Idf, linguistic Inquiry, word count

features and Dependency features. In addition, the authors conduct their experiment on the collected dataset from ASKfm website [7]. Little studies are focused on user-based feature. Therefore, in this project, we will focus on the user-based feature and we will try to extract new user-based feature to enhance the accuracy of detection cyberbullying.

Related Works:

Silva et al.[8] present first version of bully blocker application which is designed for the parents to help them monitoring Facebook interactions of their adolescent can used to detect and alerting the parents when cyberbullying occurs. Zhang et al.[9] propose a novel pronunciation based on convolution neural network to face the noise and error in social media posts and messages make detecting cyberbullying very challenging. In addition, the authors used phoneme codes of the text as features for CNN. This procedure corrects spelling errors that did not alter the pronunciation.

Romsaiyed et al.[10] present an automatic cyberbullying detection system to detect, identify, and classify cyberbullying activities from large volume of streaming text from OSN services. Text are fed into cluster and discriminate analysis stage which can identify abusive text. Then the abusive texts are clustered by using K-Mean. Naïve Bayes is used as classification method to build classifer from our training datasets and build predictive model. Dinakar et al.[11] focus on textual cyberbullying detection. In addition, the authors used different features such as tf-idf, ortony lexicon for negative affect, list of profane words, POS bigrams: jj_DT,PRP_VBP, VB_PRP and top specific unigram and bigrams. Moreover, the authors conduct their experiment on collected data from YouTube. Hee et al.[12] present automatic cyberbullying detection in social media text by modeling posts written by bullies, victims and bystanders of online bullying. Moreover, the authors describe the collection and fine-grained annotation of a training corpus for English and Dutch and perform a series of binary classification experiments. In addition, the authors extracted different features word n-gram bag-of-words, character n-grams bag-of-words, term lists, subjectivity lexicon features and topic model features. Rafiq et al.[13] design a novel approaches to detect instances of cyberbullying over Vine media sessions which is a mobile based video sharing online social network. Moreover, the authors collect a set Vine video session and us CrowdFlower to label media session for cyberbullying and cyberaggression. Rafiq et al.[14] propose a multi-stage cyberbullying detection solution that drastically reduces the classification time and the time to raise cyberbullying alerts. Moreover, the authors proposed solution which is

scalable, does not sacrifices accuracy for scalability. In addition, the solution is comprised of three novel components, an initial predictor, a multilevel priority scheduler and incremental classification mechanism. Chelmis et al.[15] perform a detailed analysis of a large-scale real world dataset to identify online social network topology structure features that are the most prominent in enhancing the accuracy of state of the art classification methods for cyberbullying detection. Moreover, the authors derived small subset of features that are fast to compute while differentiating between normal users cyberbullies and victims. Rafiq et al.[16] propose a multi-stage cyberbullying detection solution that drastically reduces the classification time and the time to raise cyberbullying alerts. Moreover, the authors proposed solution which is scalable, does not sacrifices accuracy for scalability. In addition, the solution is comprised of two novel components, dynamic priority and incremental classification mechanism. Li et al.[17] develop methods for detecting cyberbullying based on sharing images on Instagram. Moreover, the authors extracted images specific and text features from comments and from image captions. In addition, several novel features including topics determined from image captions and outputs of pretrained conventional neural network applied to image pixels. Hosseinmardi et al.[18] investigate the prediction of cyberbullying incidents in Instagram. Moreover, the authors build a predictor that can be anticipate the occurrence of cyberbullying incidents before they happen.

3. Methodology

3.1 Dataset Description

In this project, we will use the dataset on cyberbullying which is collected by Impermuim. The dataset is available on Kaggle website (https://www.kaggle.com/c/detecting-insults-in-social-commentary/data). The dataset includes the label column which contains class labels 0 and 1. Non-insult and insult respectively, followed by two features. The first feature represents that the time at which the comment was written. It is sometimes blank, meaning an accurate timestamp is not possible. It is in the form "YYYY/MM/DD/ hh mm ss" and then the Z character. It is on a 24-hour clock and corresponds to the local time at which the comment was originally written. The second feature shows that the Unicode-escaped text of the content, surrounded by double-quotes. The content is mostly English language comments, with some occasional formatting. The dataset is included 3947 instances.

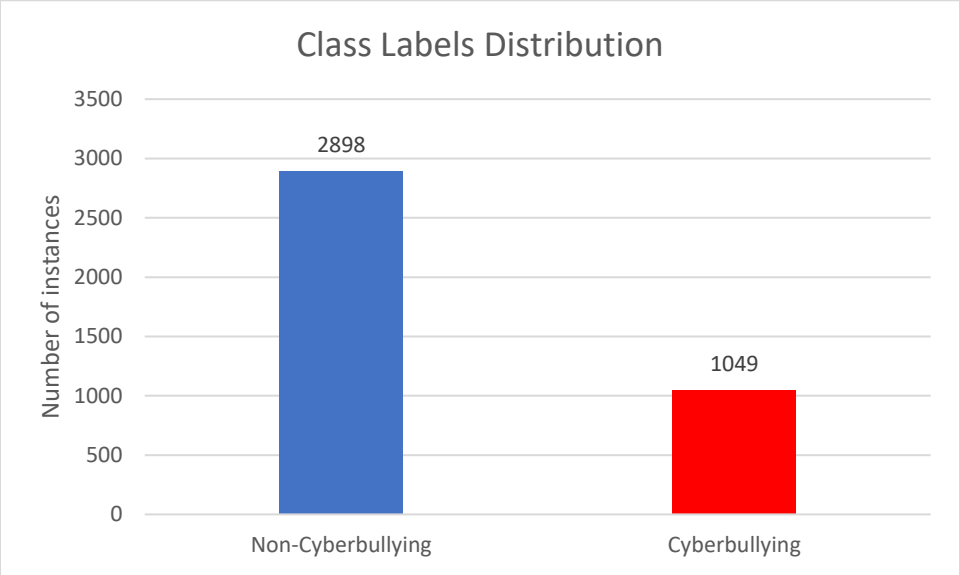

Figure 1: showing the class labels distribution in dataset

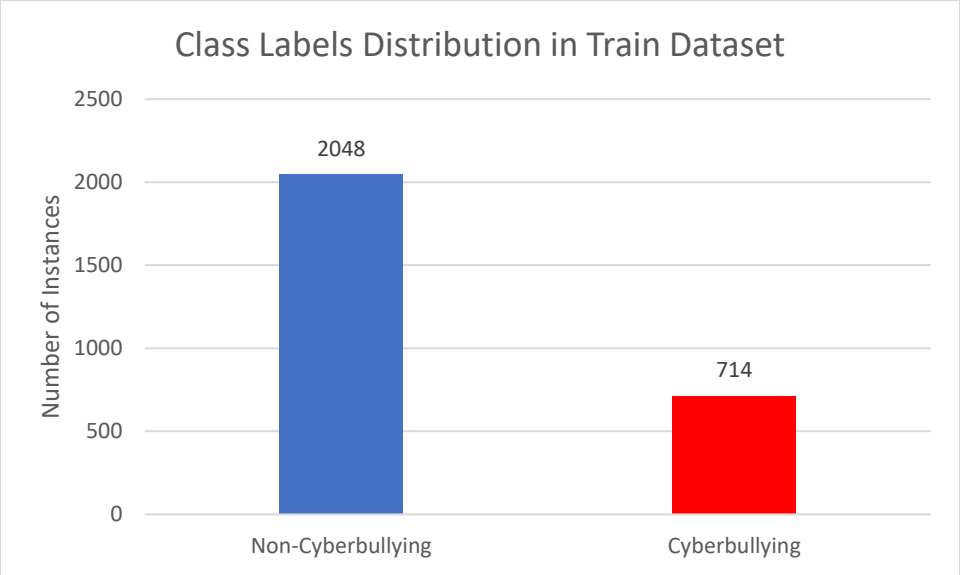

Figure 2: Showing the class labels distribution in train dataset

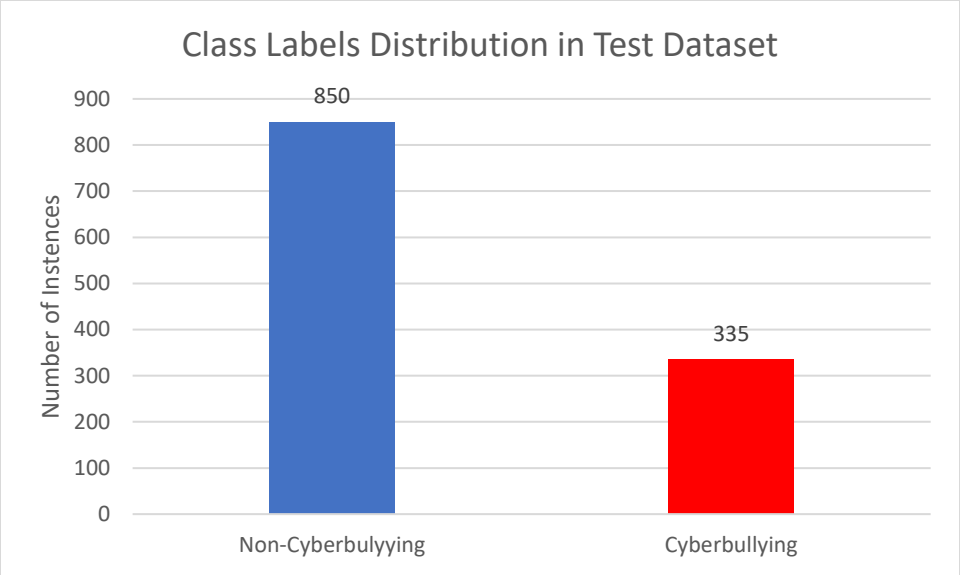

Figure 3: Showing the class labels distribution in test dataset

3.2 Textual Preprocessing and representation

A collected dataset may have noisy, redundant or incomplete data. For example, in our project, the dataset includes collected comments from YouTube website were written by people, people used to disregard the grammatical rules of writing or misspelling words. Moreover, bad data effects the knowledge discovery during training the data. Therefore, pre-processing is very important stage in knowledge discovery time line. In our project, the dataset preprocessing stage includes tokenization means that breaking the stream of text into small units called token, removing punctuations or special characters, removing stop words, converting the uppercase into lower case and stemming. Also, we apply standardization to represent the data between 0 and 1. Moreover, we split the dataset into train and test dataset. The train dataset includes 70% instances of dataset and the test dataset includes the rest.

Test representation is considered one of the problems in the data mining. Text representation means that converting unstructured text to numerically representation, so they can be computed mathematically. There are several models to represent the text such as TF-IDF, Bag of words and N-grams. In our project, we use bag of words model and N-gram model. Bag of words model is common and simply used in natural language processing application such as documents classification and clustering. Bag of words model represent the document as bag of words and measures the presence of the word in the document. Moreover, there are several methods to

compute the presence of the words such as frequency or TF-IDF. In addition, N-grams model is sequence tokens where tokens can be words, phrase or full sentences. N refers to number of sequence token. For example, “I have class tomorrow”, applying 2-gram model on the sentence so, the result will be “I have”,” have class” and” class tomorrow”.

3.3 Feature extraction for cyberbullying Detection

In our project, we use the occurrence and frequency of each word in the document as feature for training classifier. Moreover, we use the sentiment analysis as feature. sentiment analysis Sentiment Analysis is the most common text classification tool that analyses an incoming message and tells whether the underlying sentiment is positive, negative our neutral.

2.4 Machine learning methods

Machine learning is a field of computer science which concerning about the algorithms and mathematical models use to enhance the performance of computer in specific task such cyberbullying detection, fraud detection and spam detection. In our project, we use several machine learning algorithms such as perceptron, Support Vector Machine, KNN, Decision Tree and logistic Regression.

2.4 Evaluation Metrics

In our project, we use test accuracy score and cross validation.

Result and Evaluation:

In this project, we used different classification models to detect cyberbullying phenome. The tables show the test accuracy result and cross-validation accuracy result. The test accuracy is ranged 66% -75%. Also, the cross-validation accuracy result is ranged 54%-77%. In the tables 1 and 2, KNN shows high accuracy in both test accuracy and cross-validation (CV) accuracy by using Bag of words with TF-IDF. In the tables 3 and 4, SVM shows high test accuracy and CV accuracy by using 2-grams with TF-IDF. In the tables 5 and 6, KNN shows the highest test and CV accuracy by using bag of words and 2-grams with TF-IDF. In the tables 7 and 8, KNN shows the highest test and CV accuracy by using bag of words, 2-grams and sentiment analysis.

Classifiers	Test Accuracy	Running Time
Perceptron	**0.72405**	**0.04480**
SVM	**0.72995**	**0.196472**
Decision Tree	**0.69367**	**0.03896**
KNN	**0.74767**	**0.04474**
Logistic Regression	**0.72489**	**0.035907**

Table 1: Test Accuracy Results by Applying different Classifiers on test data based on Bag of words with TF_IDF presence feature

Classifiers	CV Accuracy	Running Time
Perceptron	**0.6870**	**0.498712**
SVM	**0.750182**	**2.00307**
Decision Tree	**0.667597**	**0.467095**
KNN	**0.77097**	**0.47403**
Logistic Regression	**0.74537**	**0.43911**

Table 2: CV Accuracy Results by Applying different Classifiers on test data based on Bag of words with TF_IDF presence feature

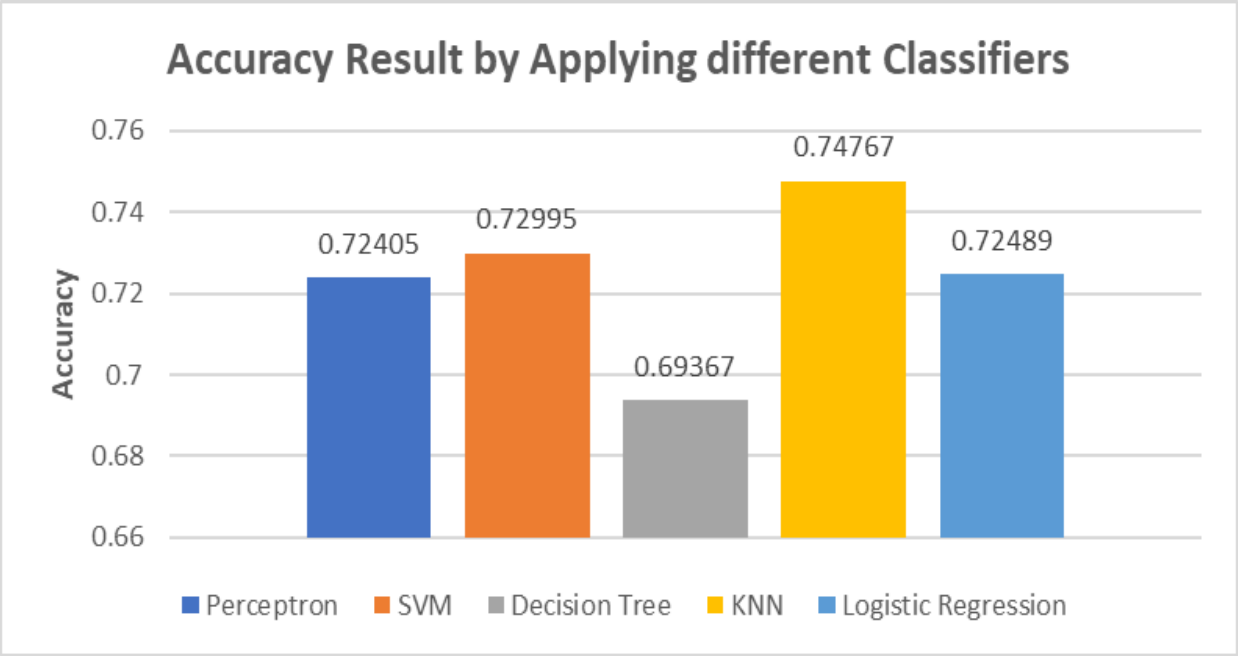

Figure4: Accuracy Results by Applying different Classifiers on test data based on bag of words with TF-IDF presence feature

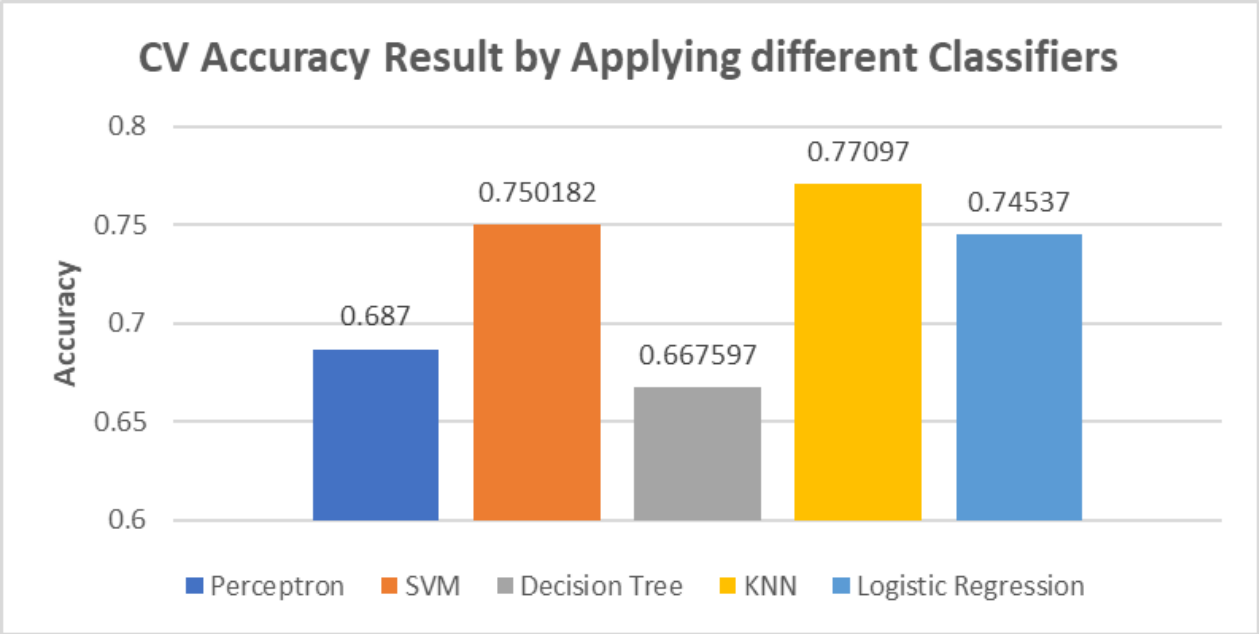

Figure 5:CV Accuracy Results by Applying different Classifiers on test data based on bag of words with TF-IDF presence feature

Classifiers	Test Accuracy	Running Time
Perceptron	**0.71814**	**0.0488**
SVM	**0.71729**	**0.04101**
Decision Tree	**0.70801**	**0.035904**
KNN	**0.71392**	**0.085791**
Logistic Regression	**0.71729**	**0.03493**

Table 3: Test Accuracy Result by Applying different Classifiers on test data based on 2-gram with TF_IDF presence feature

Classifiers	CV Accuracy	Running Time
Perceptron	**0.545378**	**0.4950**
SVM	**0.73397**	**0.72177**
Decision Tree	**0.73017**	**0.42920**
KNN	**0.72712**	**0.70405**
Logistic Regression	**0.73422**	**0.45186**

Table 4: CV Accuracy Result by Applying different Classifiers on test data based on 2-gram with TF_IDF presence feature

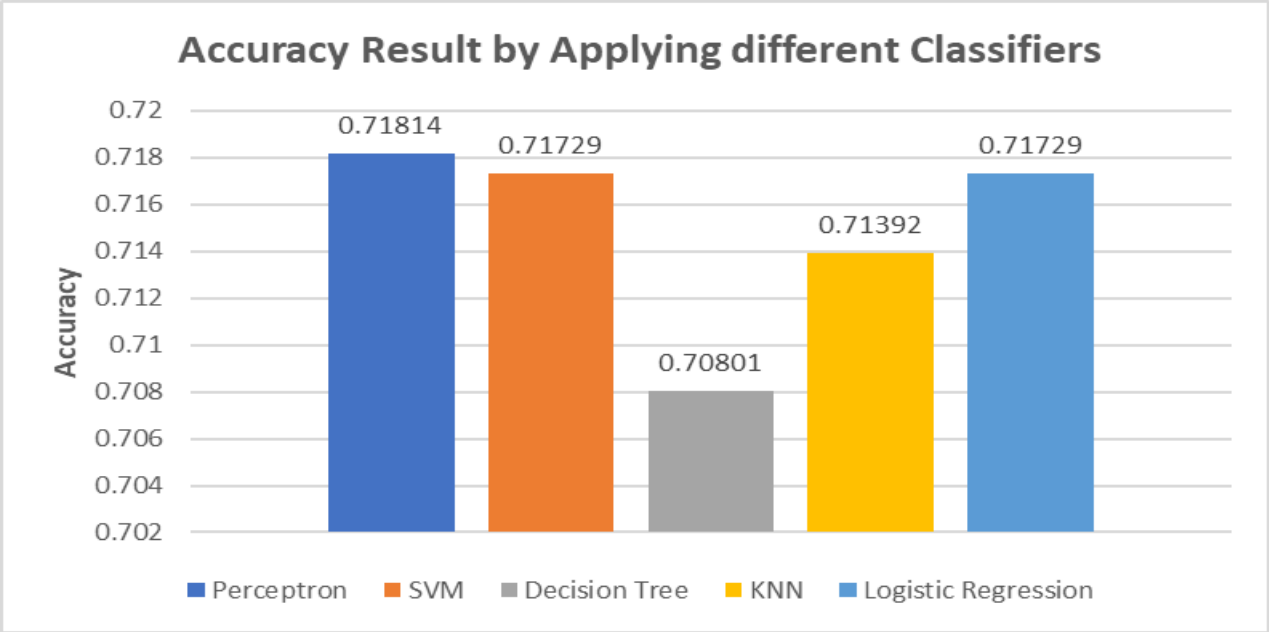

Figure 6: Test Accuracy Result by Applying different Classifiers on test data based on 2-grams with TF-IDF presence feature

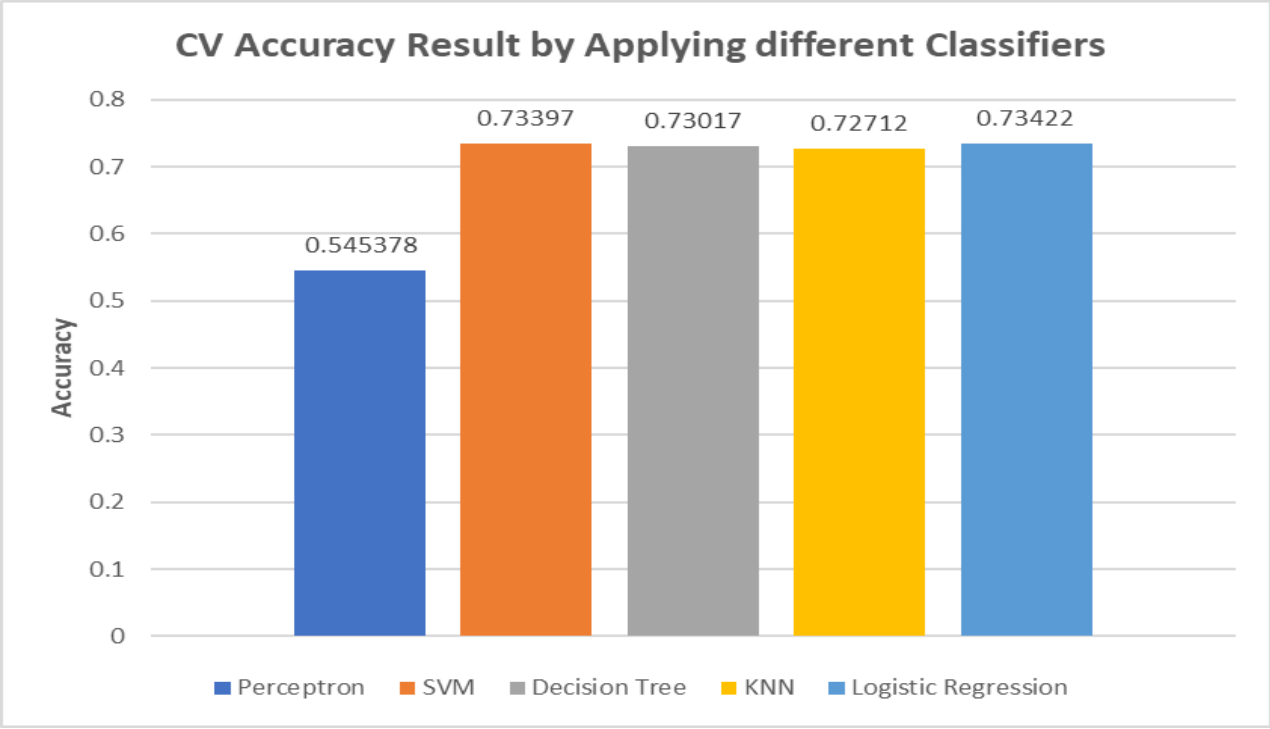

Figure 7: CV Accuracy Result by Applying different Classifiers on test data based on 2-grams with TF-IDF presence feature

Classifiers	Test Accuracy	Running Time
Perceptron	**0.66497**	**0.0628**
SVM	**0.71729957**	**0.21236**
Decision Tree	**0.694514**	**0.062830**
KNN	**0.71814**	**0.06272**
Logistic Regression	**0.717299**	**0.0550**

Table 5: Test Accuracy Result by Applying different Classifiers on test data based on 2-gram and bag of words with TF_IDF presence feature

Classifiers	CV Accuracy	Running Time
Perceptron	**0.62996**	**0.81692**
SVM	**0.73422**	**2.5048**
Decision Tree	**0.718521**	**0.7195**
KNN	**0.752985**	**0.7011**
Logistic Regression	**0.73422**	**0.7062**

Table 6: CV Accuracy Result by Applying different Classifiers on test data based on 2-gram and bag of words with TF_IDF presence feature

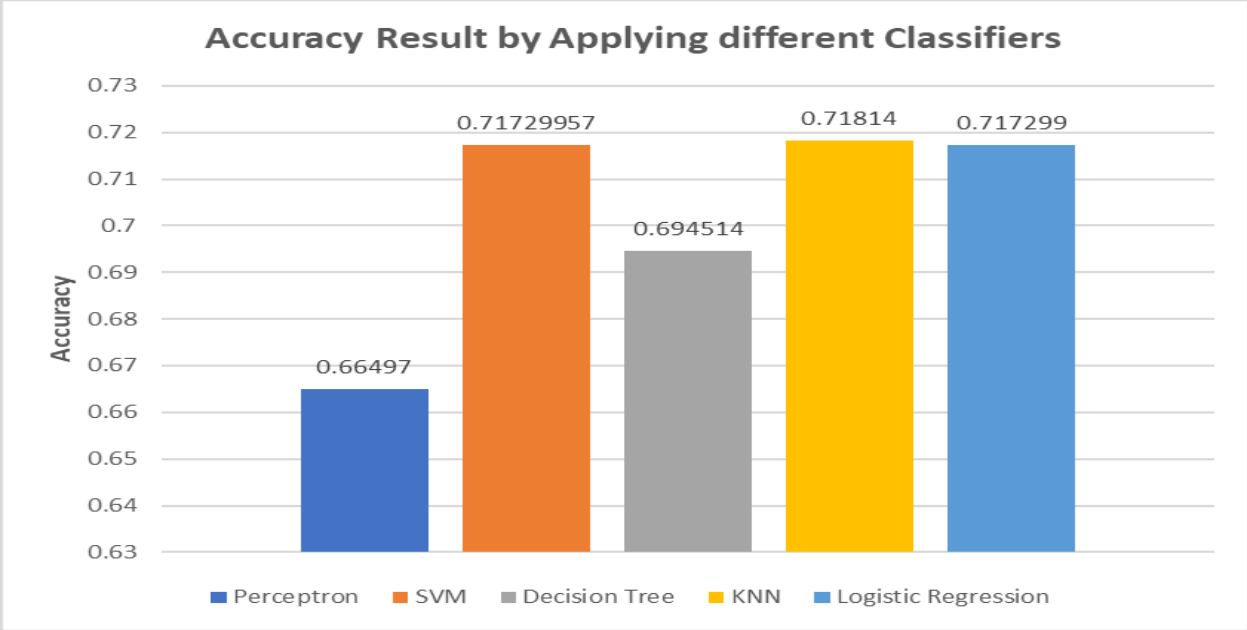

Figure 8: Test Accuracy Result by Applying different Classifiers on test data based on 2-grams and bag of words with TF-IDF presence feature

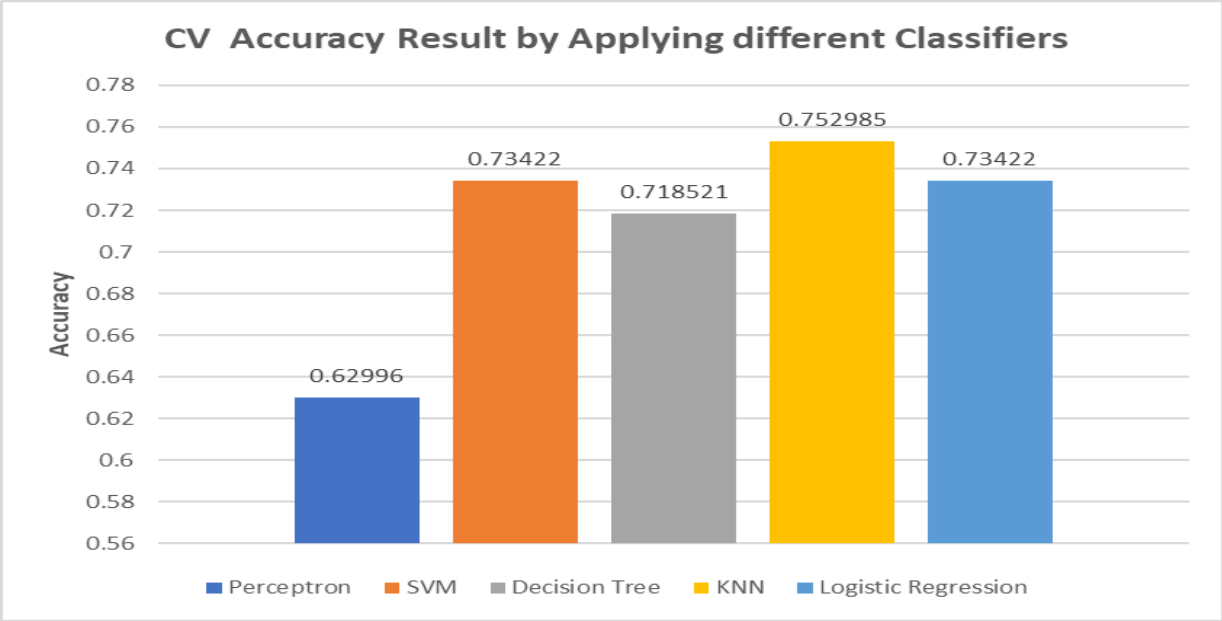

Figure 9: CV Accuracy Result by Applying different Classifiers on test data based on 2-grams and bag of words with TF-IDF presence feature

Classifiers	Test Accuracy	Running Time
Perceptron	**0.725738**	**0.06781**
SVM	**0.71729**	**0.187465**
Decision Tree	**0.72320**	**0.05585**
KNN	**0.75527**	**0.0628**
Logistic Regression	**0.717299**	**0.0550**

Table 7: Test Accuracy Result by Applying different Classifiers on test data based on 2-gram and bag of words with TF_IDF presence feature and sentiment analysis

Classifiers	CV Accuracy	Running time
Perceptron	**0.676573**	**0.77419**
SVM	**0.7344820**	**2.03125**
Decision Tree	**0.7233**	**0.711662**
KNN	**0.7600**	**0.69208**
Logistic Regression	**0.7392**	**0.69140**

Table 8: CV Accuracy Result by Applying different Classifiers on test data based on 2-gram and bag of words with TF_IDF presence feature and sentiment analysis

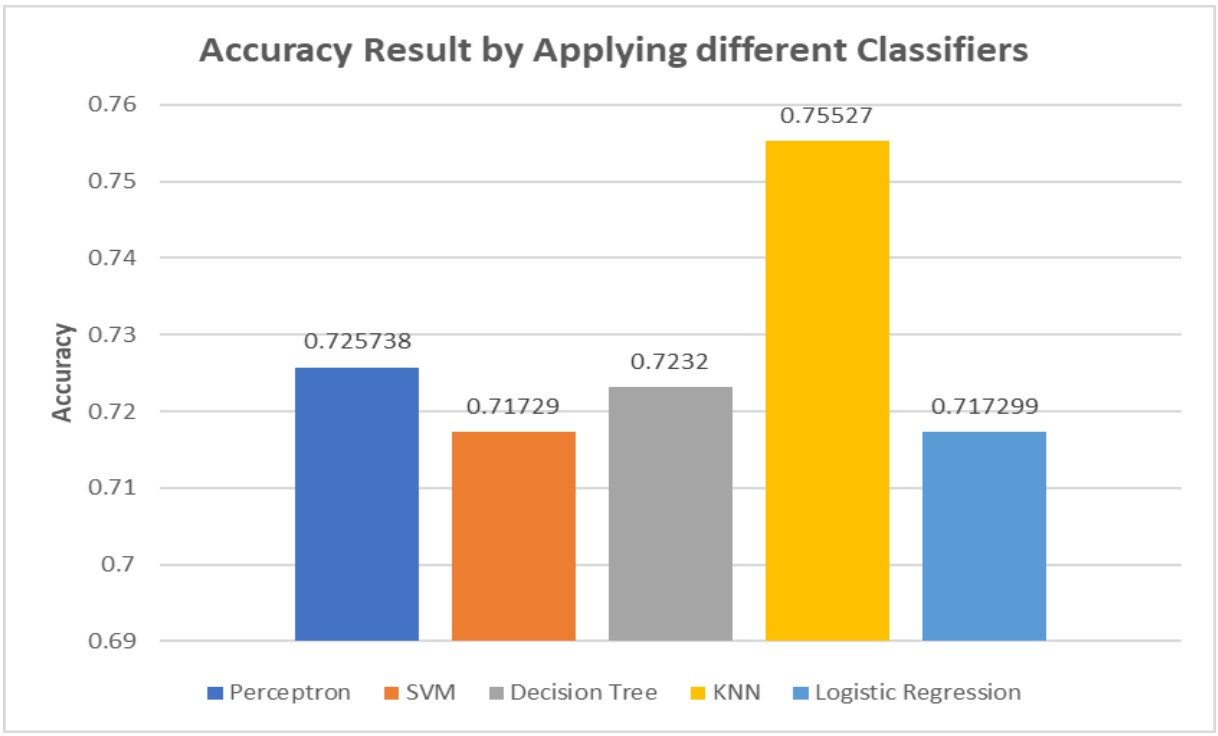

Figure 10: Test Accuracy Result by Applying different Classifiers on test data based on 2-gram and bag of words with TF_IDF presence feature and sentiment analysis

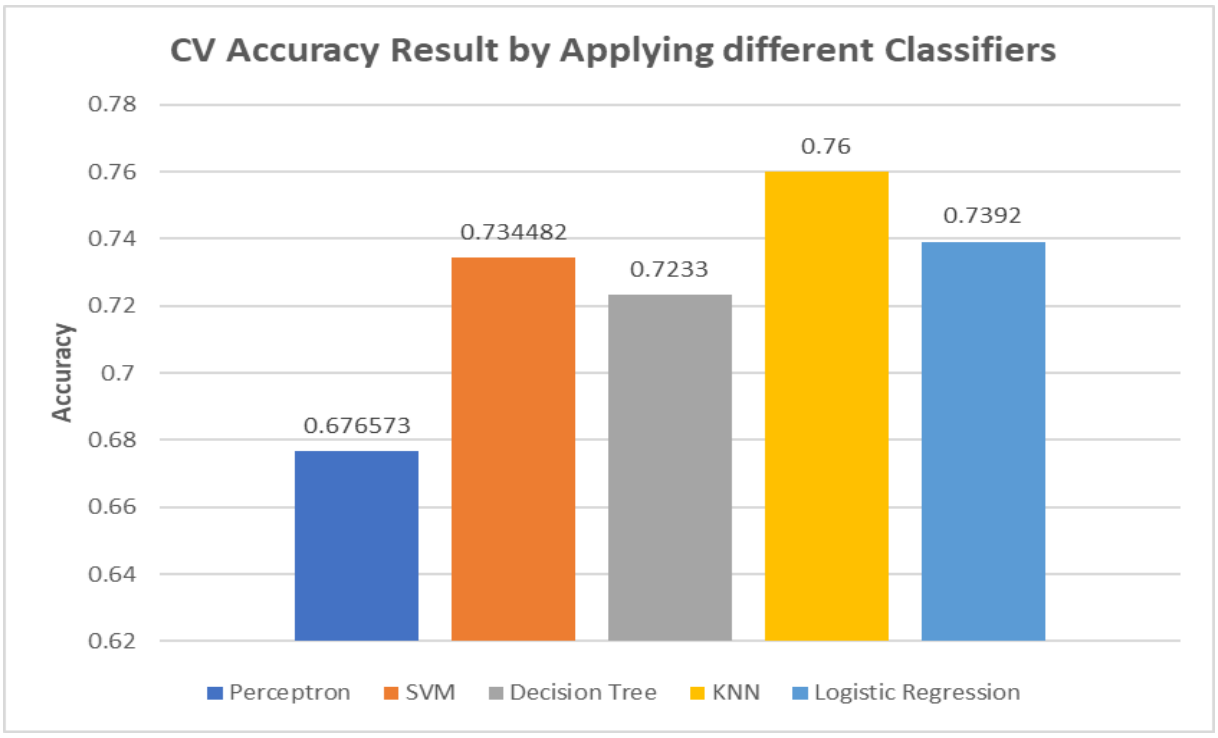

Figure 11: Test Accuracy Result by Applying different Classifiers on test data based on 2-gram and bag of words with TF_IDF presence feature and sentiment analysis

References:

[1] “What is Cyberbullying” Accessed March 15,2018. [online]. Available: https://www.stopbullying.gov/cyberbullying/what-is-it/index.html

[2] National Center for Education Statistics and Bureau of Justice *Statistics*, *2011*.

[3] Kosciw, J. G., Greytak, E. A., Bartkiewicz, M. J., Boesen, M. J., & Palmer, N. A, “The 2011 National School Climate Survey: The experiences of lesbian, gay, bisexual and transgender youth in our nation’s schools*”. New York: GLSEN.*

[4] Sourander, A., Brunstein-Klomek, A., Ikonen, M., Lindroos, J., Luntamo, T., Koskelainen, M., Ristkari, T., Hans Helenius, H. 2010. “Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study.” *Arch Gen Psychiatry*, 67: 720-728.

[5] S. Hinduja and J. W. Patchin. “Bullying, cyberbullying, and suicide”. *Archives of suicide research*, 14(3):206–221, 2010.

[6] Robin M Kowalski and Susan P Limber.” Electronic bullying among middle school students”. *Journal of adolescent health*, 2007,41(6):S22–S30.

[7] Y. J. Foong and M. Oussalah, "Cyberbullying System Detection and Analysis," *2017 European Intelligence and Security Informatics Conference (EISIC)*, Athens, 2017, pp. 40-46.
doi: 10.1109/EISIC.2017.43

[8] Y. N. Silva, C. Rich, J. Chon and L. M. Tsosie, "BullyBlocker: An app to identify cyberbullying in facebook," *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, San Francisco, CA, 2016, pp. 1401-1405. doi: 10.1109/ASONAM.2016.7752430

[9] X. Zhang *et al*., "Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network," *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Anaheim, CA, 2016, pp. 740-745. doi: 10.1109/ICMLA.2016.0132

[10] W. Romsaiyud, K. na Nakornphanom, P. Prasertsilp, P. Nurarak and P. Konglerd, "Automated cyberbullying detection using clustering appearance patterns," *2017 9th International Conference on Knowledge and Smart Technology (KST)*, Chonburi, 2017, pp. 242-247. doi: 10.1109/KST.2017.7886127

[11] K. Dinakar, R. Reichart, and H. Lieberman, “Modeling the Detection of Textual Cyberbullying,” in Proc. IEEE International Fifth International AAAI Conference on Weblogs and Social Media (SWM’11), Barcelona, Spain, 2011.

[12] Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W. and Hoste, V. (2018). Automatic Detection of Cyberbullying in Social Media Text. [online] Arxiv.org. Available at: https://arxiv.org/abs/1801.05617 [Accessed 11 May 2018].

[13] R. Ibn Rafiq, H. Hosseinmardi, R. Han, Qin Lv, S. Mishra and S. A. Mattson, "Careful what you share in six seconds: Detecting cyberbullying instances in Vine," *2015 IEEE/ACM International Conference on*

Advances in Social Networks Analysis and Mining (ASONAM), Paris, 2015, pp. 617-622. doi: 10.1145/2808797.2809381

[14]. R. Ibn Rafiq, H. Hosseinmardi, R. Han, Qin Lv, S. Mishra, “Investigating Factors Influencing the Latency of Cyberbullying Detection”.

[15] C. Chelmis, D. S. Zois and M. Yao, "Mining Patterns of Cyberbullying on Twitter," *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, New Orleans, LA, 2017, pp. 126-133. doi: 10.1109/ICDMW.2017.22

[17] Zhong, H., Li, H., Squicciarini, A., Rajtmajer, S., Griffin, C., Miller, D., & Caragea, C. *“*Content-Driven Detection of Cyberbullying on the Instagram Social Network". in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, 2016.

[18] H. Hosseinmardi, R. I. Rafiq, R. Han, Q. Lv and S. Mishra, "Prediction of cyberbullying incidents in a media-based social network," *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, San Francisco, CA, 2016, pp. 186-192. doi: 10.1109/ASONAM.2016.7752233